Questioning the rater idiosyncrasy explanation for error variance by searching for multiple signals within the noise

Addendum
Valorization paragraph

1. *(Relevance) What is the social (and/or economic) relevance of your research results (i.e. in addition to the scientific relevance)?*

Variability in ratings, human judgment, assessment and measurement were discussed within this dissertation. Rater-based assessments are an important component of programmatic assessment[1] spanning all levels of medical training and requiring significant financial and human resources. However, the quality of our assessment systems affects more than just our medical training programs. Since training per student is expensive in terms of money and time, many medical schools are heavily funded through taxpayer money. As such, governments demand accountability from the educational institutions for those funds and taxpayers expect quality health care as a result of the expenditures. Assessment outcomes are one form of evidence medical training programs can use in their reporting their activities.

More critically, assessment functions as a gate-keeping mechanism to determine when students are sufficiently competent to progress to the next phase of training. This includes assessment being used to determine when trainees are competent enough to be considered autonomous practitioners who provide unsupervised health care to the general public. As emphasized by Kogan and colleagues, "To be professionally accountable and attain the public's trust, the onus is on us, as medical educators, to make good assessment decisions. The interrater variability of work-based assessments is not just an educational issue but also a patient care and safety issue. Medical education and health care delivery are intertwined."[2(p.725)] Our failure to adequately assess students and trainees could negatively impact health care delivery making assessment design and analysis relevant to societal needs.

This is especially important since there has been a world-wide shift to adopting outcomes-based medical education at all levels of training.[3] The outcomes to be measured are complex, functionally relevant skills; skills which are vital for trainees to perform well within the labour market.[4] Assessment of these complex skills requires human judgment, and therefore, understanding how humans make assessment judgments is important. If we fail to use human judgment well within our assessment programs, outcomes-based medical education will fail. This dissertation prompts further investigations into better understanding human judgment within assessment tasks.

2. *(Target groups) To whom, in addition to the academic community, are your research results of interest and why?*

The general public relies on our graduates to provide health care. They expect our assessments to determine if and when trainees are capable of providing quality care.[5] As such, medical licensing examination boards may be interested in considering these

results when making decisions regarding which measurement techniques to use. The need for National licensing examinations is debated[6] and their role might change if more sound workplace-based assessments could be used to map the progression of trainee competence over time.

Additional target groups include program administrators who could use these results to re-consider how they collect and interpret assessments from supervisors. For example, these results suggest that administrators should expect multiple different assessment judgments to be collected from different supervisors for the same trainee. These different judgments are associated with different points of view but are not likely to be unique to each supervisor and do not necessarily reflect bias in the judgments. Similarly, clinical mentors could consider these results when meeting with trainees to discuss their clinical performance. Discussions could include reflections about why performance can be perceived differently than it was intended. Trainees could use these results to prepare themselves to receive differing judgments of their performance. Rather than dismissing them as mixed messages, these results might help encourage trainees to view different judgments as differing critiques.

3. *(Activities/Products) Into which concrete products, services, processes, activities or commercial activities will your results be translated and shaped?*

I believe it is premature to begin translating these results into assessment products. It is reassuring that a relatively limited number of variations between raters were found rather than idiosyncratic impressions and judgments. Although, these results do have some unsettling implications for how we conceptualize our current assessment designs *and* radical design changes would be required to accommodate them. However, these results are very preliminary and need to be extended and challenged before it would be reasonable to use them to inform design changes. I would be delighted if fellow researchers were motivated to test the limits of these findings and through their investigations, along with our own, we could determine the robustness of the effects.

4. *(Innovation) To what degree can your results be called innovative in respect to the existing range of products, services, processes, activities and commercial activities?*

This dissertation is innovative in presenting a new way to think about variability in ratings. It takes a novel approach to trying to understand rater cognition by assuming most raters are trying to provide useful assessment judgments and are capable of providing quality assessment judgments. Up until this point, researchers in favour of using ratings for performance assessments have viewed variability as something to minimize (as is discussed in more detail in Chapter 2). Researchers who have been

more critical of using ratings for performance assessment have focused on the (mis)alignment of the psychometric assumptions to the purpose of the assessments.[7,8] The focus has been at the conceptual or philosophical level.[9] This program of research provides some initial evidence that physicians may be not interchangeable as raters, and therefore, violate the psychometric assumption for homogeneity of the rater population. These results could be used to challenge the use of psychometric measurement for performance assessment ratings, calling for new views on how to handle assessment information.[10]

These results are innovative in that they compel us to start thinking differently about the source of variability in ratings and what the variability reveals about rater cognition. If these results are found to be robust through further investigations, they could inspire a novel approach to assessment design and analysis. To be consistent with this dissertation's perspective, the novel approach could include first determining what assessment information supervisors can aptly provide and then designing a suitable assessment system to collect and analyze it. As an example of one possible direction for innovation in assessment design, we presented a hypothetical assessment design that does not use ratings. This dissertation changes what is seen as the problem with rater-based assessments, and subsequently, opens up a new potential for solutions.

5. *(Schedule & Implementation) How will this/these plans(s) for valorization be shaped? What is the schedule, are there risks involved, what market opportunities are there and what are the costs involved?*

For the last several years, I have been discussing these questions and findings at academic conferences and asking other researchers for their reactions to them. This has been immensely helpful in refining and re-directing my thinking around rater cognition. It has also led to invitations to join research collaborations that will hopefully allow me to pursue multiple lines of research into assessment judgments, in parallel. As I transition from being a research associate to a faculty member, I believe I am well-positioned to continue investigating variability in assessment judgments. This research will aim to contribute to the medical education community's current interests in using entrustable professional activities and competency-based assessments. Rater-based assessments are a key component of our apprenticeship-like clinical training models resulting in a strong need for improved assessment designs. Even though it is too early to specify when and how it will happen, as more is uncovered about rater cognition, I am confident the new findings will contribute to improving assessment designs and decisions regarding trainee competency.

REFERENCES

1 Van der Vleuten C, Schuwirth L, Driessen E, Dijkstra J, Tigelaar D, Baartman L et al. A model for programmatic assessment fit for purpose. *Medical Teacher* 2012;34 (3):205-14.

2 Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Academic Medicine* 2014;89 (5):721-7.

3 Tekian A, Hodges BD, Roberts TE, Schuwirth L, Norcini J. Assessing competencies using milestones along the way. *Medical Teacher* 2015;37 (4):399-402.

4 Ben-David MF. The role of assessment in expanding professional horizons. *Medical Teacher* 2000;22 (5):472-7.

5 Otaki J, Nagata-Kobayashi S, Takayashiki A. Aspects of clinical skills test demanded by the public for the national medical licensure examination in japan. *Medical Teacher* 2012;34 (5):423-.

6 Bajammal S, Zaini R, Abuznadah W, Al-Rukban M, Aly SM, Boker A et al. The need for national medical licensing examination in saudi arabia. *BMC Medical Education* 2008;8:53-.

7 Schuwirth LW, van der Vleuten CP. Assessing competence. In: Hodges BD, Lingard LA, editors. The question of competence: Reconsidering medical education in the twenty-first century. First Ed. Ithaca and London: ILR Press-Cornell University Press; 2012. p. 113-30.

8 Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher* 2013;35 (7):564-8.

9 Whitehead CR, Kuper A, Hodges B, Ellaway R. Conceptual and practical challenges in the assessment of physician competencies. *Medical Teacher* 2015;37 (3):245-51.

10 Govaerts M, Vleuten CPM. Validity in work-based assessment: Expanding our horizons. *Medical Education* 2013;47 (12):1164-74.